In the Claims:

Claims 1, 28, and 36 are amended.

Claim 9 is currently canceled.

Claims 1, 3-6, 10-19, and 28-39 are pending.

1. (Currently amended) A method of using a tuning set of information to jointly optimize the performance and size of a language model, comprising:

segmenting at least a subset of a received textual corpus into segments by clustering every N-items of the received corpus into a training unit, wherein resultant training units are separated by gaps, and wherein N is an empirically derived value based, at least in part, on the size of the received corpus;

creating the tuning set from application-specific information;

(a) training a seed model via the tuning set;

(b) calculating a similarity within a sequence of the training units on either side of each of the gaps;

(c) selecting segment boundaries that maximize intra-segment similarity and inter-segment disparity;

(d) calculating a perplexity value for each segment based on a comparison with the seed model;

(e) selecting some of the segments based on their respective perplexity values to augment the tuning set;

iteratively refining the tuning set and the seed model by repeating steps (a) through (e) until a threshold;

and

refining the language model based on the seed model.

2. (Canceled)

3. (Original) A method according to claim 1, wherein the tuning set of information is comprised of one or more application-specific documents.

4. (Original) A method according to claim 1, wherein the tuning set of information is a highly accurate set of textual information linguistically relevant to, but not taken from, the received textual corpus.

5. (Original) A method according to claim 1, further comprising a training set comprised of at least the subset of the received textual corpus.

6. (Original) A method according to claim 5, further comprising:

ranking the segments of the training set based, at least in part, on the calculated perplexity value for each segment.

7. (Canceled)

8. (Canceled)

9. (Canceled)

10. (Previously presented) A method according to claim 1, wherein the calculation of the similarity within a sequence of training units defines a cohesion score.

11. (Original) A method according to claim 10, wherein intra-segment similarity is measured by the cohesion score.

12. (Previously presented) A method according to claim 10, wherein inter-segment disparity is approximated from the cohesion score.

13. (Original) A method according to claim 1, wherein the calculation of inter-segment disparity defines a depth score.

14. (Original) A method according to claim 1, wherein the perplexity value is a measure of the predictive power of a certain language model to a segment of the received corpus.

15. (Original) A method according to claim 1, further comprising:

ranking the segments of at least the subset of the received corpus based, at least in part, on the calculated perplexity value of each segment; and

updating the tuning set of information with one or more of the segments from at least the subset of the received corpus.

16. (Original) A method according to claim 15, wherein one or more of the segments with the lowest perplexity value from at least the subset of the received corpus are added to the tuning set.

17. (Original) A method according to claim 1, further comprising:

utilizing the refined language model in an application to predict a likelihood of another corpus.

18. (Original) A storage medium comprising a plurality of executable instructions including at least a subset of which, when executed, implement a method according to claim 1.

19. (Original) A system comprising:

a storage medium having stored therein a plurality of executable instructions; and

an execution unit, coupled to the storage medium, to execute at least a subset of the plurality of executable instructions to implement a method according to claim 1.

20-27. (Canceled)

28. (Currently amended) A modeling agent comprising:

a controller, to receive invocation requests to develop a language model from a corpus; and

a data structure generator, responsive to the controller, to:

develop a seed model from a tuning set of information;

segment at least a subset of a received corpus, wherein the segments of the received corpus are a clustering of every N items of the received corpus into a training unit, wherein N is an empirically derived value based, at least in part, on the size of the received corpus, and the training units are separated by gaps;

calculate the similarity within a sequence of training units on either side of each of the gaps;

select segment boundaries that improve intra-segment similarity and inter-segment disparity;

calculate a perplexity value for each segment;

refine the seed model with one or more segments of the received corpus based, at least in part, on the calculated perplexity values;

iteratively refine the tuning set with segments ranked by the seed model and in turn iteratively update the seed model via the refined tuning set;

filter the received corpus via the seed model to find low-perplexity segments; and

train the language model via the low-perplexity segments.

29. (Original) A modeling agent according to claim 28, wherein the tuning set is dynamically selected as relevant to the received corpus.

30. (Original) A modeling agent according to claim 28, the data structure generator comprising:

a dynamic lexicon generation function, to develop an initial lexicon from the tuning set, and to update the lexicon with select segments from the received corpus.

31. (Original) A modeling agent according to claim 28, the data structure generator comprising:

a frequency analysis function, to determine a frequency of occurrence of segments within the received corpus.

32. (Original) A modeling agent according to claim 28, the data structure generator comprising:

a dynamic segmentation function, to iteratively segment the received corpus to improve a predictive performance attribute of the modeling agent.

33. (Original) A modeling agent according to claim 32, wherein the dynamic segmentation function iteratively re-segments the received corpus until the language model reaches an acceptable threshold.

34. (Original) A modeling agent according to claim 32, the data structure generator further comprising:

a frequency analysis function, to determine a frequency of occurrence of segments within the received corpus.

35. (Original) A modeling agent according to claim 34, wherein the data structure generator selectively removes segments from the data structure that

do not meet a minimum frequency threshold, and dynamically re-segments the received corpus to improve predictive capability while reducing the size of the data structure.

36. (Currently Amended) A method of jointly optimizing the performance and size of a language model, comprising:

segmenting one or more relatively large language corpora into multiple segments of N items, wherein N is an empirically derived value based, at least in part, on the size of the received corpus; ~~equal size~~

selecting an initial tuning sample of application-specific data, the initial tuning sample being relatively small in comparison to the one or more relatively large language corpora, wherein the initial tuning sample is used for training a seed model, the seed model to be used for ranking the multiple segments from the language corpora;

iteratively training the seed model to obtain a mature seed model, wherein the iterative training proceeds until a threshold is reached, each iteration of the training including:

updating the seed model according to the tuning sample;

ranking each of the multiple segments according to a perplexity comparison with the seed model;

selecting some of the multiple segments that possess a low perplexity; and

augmenting the tuning sample with the selected segments;

once the threshold is reached, filtering the language corpora according to the mature seed model to select low-perplexity segments;

combining data from the low-perplexity segments; and training the language model according to the combined data.

37. (Previously Presented) The method as recited in claim 36, wherein the selecting an initial tuning sample comprises selecting a few application-specific documents.

38. (Previously Presented) The method as recited in claim 36, wherein the threshold comprises one of a predetermined size of the seed model or a sufficient application specificity of the seed model.

39. (Previously Presented) The method as recited in claim 36, further comprising pruning the language model utilizing an entropy based cutoff algorithm that uses only information embedded in the language model itself.